УДК 519:616-079.4:616.5

Е.В. ВЫСОЦКАЯ**,** канд. техн. наук, доц. ХНУРЭ (г. Харьков),
А.Н. БЕЛОВОЛ**,** доц. ХНМУ (г. Харьков),
Ю.В. КИРИЧЕНКО (г. Харьков)

ВОССТАНОВЛЕНИЕ ПРОПУЩЕННЫХ ЗНАЧЕНИЙ ПАРАМЕТРОВ В ТАБЛИЦАХ БИОХИМИЧЕСКИХ АНАЛИЗОВ ПАЦИЕНТОВ С ПСОРИАЗОМ

В статье рассмотрен подход к восстановлению пропущенных значений с помощью нейросетевых технологий. Предложена нейронная сеть, позволяющая восстанавливать отсутствующие клинико-биохимические показатели пациентов с псориазом. Проведен расчет, подставляемых на место пропуска значений, и оценена адекватность восстановления данных на искусственно введенных пропусках. Ил.: 1. Табл. 3. Библиогр.: 9 назв.

Ключевые слова: восстановление пропущенных значений, нейросетевая технология, клинико-биохимические показатели, псориаз.

Постановка проблемы и анализ литературы. Псориаз – один из самых распространенных хронических дерматозов, которым страдает от 1 до 5% населения планеты. В последнее время все чаще о псориазе говорят как о системном заболевании из-за вовлечения в процесс не только кожи, но и суставов, ряда внутренних органов [1]. Поэтому для выяснения причин возникновения псориаза необходимо учитывать состояние всех органов и систем организма в целом. Но для установления состояния каждой из систем необходимо взять у пациента множество различных анализов, что представляет некоторую сложность. Поэтому проблема обработки и анализа информации с пропусками биохимических параметров пациентов и ее дальнейшего использования для выяснения причин появления псориаза является актуальной. Анализируемая выборка должна отвечать критериям качественности и полноты. В реальности приходится сталкиваться с ситуацией, когда некоторые из свойств одного или нескольких объектов отсутствуют – возникает ситуация данных с пропусками, что значительно осложняет математическую обработку, так как смещение основных статистических характеристик, таких как математическое ожидание или дисперсия, возрастает, например, прямо пропорционально числу пропусков. Поэтому проблема предварительной обработки данных является одной из первостепенных. Основываясь на тех или иных представлениях о природе пропущенных значений, имеются различные способы их заполнения.

Существует несколько подходов к анализу медицинских данных с пропусками [2]. К ним относятся следующие: удаление всего комплекта, если он содержит хотя бы один пропуск; замена пропуска на условное значение, например, null, с дальнейшим пропуском при обработке; дополнение пропущенных значений.

Рассмотрим варианты работы с комплексом данных, содержащих пропущенные значения.

Самым простым решением задачи обработки результатов исследования является исключение некомплектных наблюдений, содержащих пропуски, и дальнейший анализ полученных таким образом "полных" данных. Понятно, что такой подход приводит к сильному различию статистических выводов, сделанных при наличии в данных пропусков и при их отсутствии.

Поэтому более перспективным является иной путь – заполнение пропусков перед анализом фактических значений. Можно выделить следующие преимущества данного подхода: ясное представление структуры данных; вычисление необходимых итоговых значений; уверенная интерпретация результатов анализа, что позволяет опираться на традиционные характеристики и суммарные значения [3].

Все существующие алгоритмы заполнения пропусков в данных можно разделить на два больших класса: простые алгоритмы и сложные алгоритмы.

Простые алгоритмы – неитеративные алгоритмы, основанные на простых арифметических операциях, расстояниях между объектами, регрессионном моделировании. К ним относится заполнение пропусков средним арифметическим, регрессионное моделирование пропусков, метод HotDeck и подбор в группе [4].

В результате применения метода заполнения пропусков средними значениями, несколько значений одного фактора оказываются одинаковыми, что указывает на его низкую точность.

В методе ближайших соседей находят строки таблицы, которые по определенному критерию (обычно, минимума декартового расстояния), являются ближайшими к строке с пропуском.

Для его заполнения значения фактора у соседей усредняются с весовыми коэффициентами, обратно пропорциональными их декартовому расстоянию к строке, которая содержит пропуск. Метод точнее предыдущего, но он практически неприменим в случае большого количества пропусков и базируется на предположении о существовании связей между объектами.

В регрессионном методе по комплектным данным строится уравнение линейной множественной регрессии, и вычисляются пропущенные значения факторов. Метод невозможно применить, если количество пропусков в строке больше одного, что приводит к множеству решений, и кроме того, в реальных задачах зависимости, чаще всего, нелинейные, поэтому его точность является невысокой.

Сложные алгоритмы – итеративные алгоритмы, предполагающие оптимизацию некоторого функционала, отражающего точность расчета подставляемых на место пропуска значений. Их можно разделить на глобальные и локальные.

Локальные алгоритмы – в оценивании (предсказании) каждого пропущенного значения участвуют полные наблюдения, находящиеся в

некоторой окрестности предсказываемого объекта. К данной группе относятся алгоритмы Zet и Zet Braid.

Главная идея алгоритма ZET заключается в циклическом формировании "компетентной матрицы", подборе параметров модели прогнозирования и прогнозировании пропуска. Недостатком алгоритма является его локальность, поскольку для вычисления отсутствующего значения используются не все данные таблицы, а лишь их часть. Субъективизм определения размерности "компетентной матрицы" приводит к учету неинформативных "шумовых" факторов и смещению оценки неизвестного значения.

Основное отличие алгоритма ZetBraid от алгоритма ZET заключается в формировании "компетентной матрицы". В процессе работы алгоритма происходит последовательный поочередный отбор компетентных строк и компетентных столбцов. Критерием оценки адекватности компетентной матрицы выступает оценка качества предсказания неизвестного элемента. Все другие недостатки, в том числе и статистическая оценка неизвестного значения исключительно на основе корреляционно-регрессионного анализа, остаются.

Глобальные алгоритмы – в оценивании (предсказании) каждого пропущенного значения участвуют все объекты рассматриваемой совокупности: метод Бартлетта, EM-оценивание и Resampling и другие.

Метод максимальной правдоподобности и EM-алгоритм требует проверки гипотез о распределении значений факторов. Применение осложняется в случае большого количества пропущенных значений фактора.

Метод Бартлетта применяется для заполнения пропусков в векторе значений результирующей характеристики в допущении, что значения входных факторов являются комплектными. Его недостатком является базирование на предположении о линейной зависимости, но отсутствие обоснования применимости метода наименьших квадратов приводит к ошибкам.

Метод Resampling имеет те же недостатки, что и предыдущий. Он является итеративным и имеет две модификации. В первой из них некомплектные строки случайным образом заменяют на комплектные из исходной матрицы и рассчитывают уравнение регрессии. Во втором варианте уравнение регрессии получают из комплектной подматрицы, находят оценки неизвестных значений, ищут уравнение регрессии. После определенного количества итераций значения коэффициентов усредняют. Информационная избыточность на фоне малой мощности множества комплектных данных в первой модификации resampling и информационная недостаточность в композиции со случайным формированием значений исходной характеристики не позволяют получать приемлемые результаты. Кроме того, отсутствуют процедуры оптимизации метода.

Рассматривая вышеперечисленные методы, делаем вывод об их низкой точности, наличии жестких требований к исходной информации, количеству

пропусков, размерности матрицы данных, априорных предположениях о существующих зависимостях, сложности реализации, что свидетельствует о необходимости разработки методов, базирующихся на новых подходах, таких как нейронные сети.

Нейронные сети могут обучаться любым функциям, что позволяет избежать использования сложного математического аппарата. Использование нелинейных функций активации позволяет решать задачи с нелинейностями.

Целью данной работы является разработка метода восстановления пропущенных значений параметров в таблицах биохимических анализов пациентов с псориазом на основе нейронной сети.

Постановка задачи восстановления пропусков в биохимических показателях. Рассмотрены показатели белкового (ast, alt, urea и т.д.), углеводного (gluc, Mg, Г-6ФФГ и т.д.), микроэлементного (Fe, Ca, Mg и т.д.) и жирового (ИБХЛ, МДА, Диены и т.д.) обменов. Всего рассмотрено 46 различных показателей, которые были взяты у 153 пациентов с псориазом. У 88 человек значения по всем показателям были заполнены полностью, а у 65 пациентов комплект был заполнен на 95 – 99%.

Для обучения нейронной сети были взяты данные тех пациентов, которые содержали полный комплект значений биохимических показателей. Для восстановления пропущенных значений использовалась уже обученная нейронная сеть.

Пусть $X = \{X_1, X_2, ..., X_n\}$, $n = \overline{1,\ 153}$ – вектор входных биохимических показателей, Y – вектор диагнозов, значения элементов которого соответствуют различным формам псориаза (распространенный псориаз, артропатическая форма псориаза и т.д.), m – размерность каждого входного вектора (m = 46). Исходная информация представлена в табл. 1. Она имеет пропуски, обозначенные звездочками.

Таблица 1

Таблица исходных данных биохимических показателей с пропусками

№ исследования	**Параметры биохимических показателей**					**Диагноз**
	1	**2**	**3**	**…**	***m***	***Y***
1	X_{11}	X_{12}	X_{13}	…	X_{1m}	Y_1
2	X_{21}	X_{22}	*	…	X_{2m}	Y_2
3	X_{31}	X_{32}	X_{33}	…	*	Y_3
…	…	…	…	…	…	…
n	X_{n1}	*	X_{n3}	…	X_{nm}	Y_n

При решении задачи восстановления пропущенных значений минимизируется функция:

$$f = \arg\min_{*} \left| Y - F(X) \right|,$$

где, $F = F(X_1, X_2, ..., X_m)$ – функция, определяющая взаимосвязь между выходной переменной Y и вектором X входных переменных

$$Y_i = F_i(X_{i1}, X_{i2}, ..., X_{im}), \quad i = \overline{1, n} . \qquad (1)$$

Поэтому задача восстановления пропущенных значений сводится к определению соотношений (1).

Решение задачи восстановления данных в биохимических показателях. Существуют множество видов структур нейронных сетей, каждая из которых предназначена для решения определенных типов задач. Применение GRNN сети для решения задачи по восстановлению пропусков данных обусловлено следующими ее преимуществами [4]:

– возможность моделирования нелинейных связей между входными и выходными параметрами;

– архитектура сети фиксирована и не нуждается в определении;

– время обучения сети значительно меньше, чем у других ИНС.

Создадим искусственную нейронную сеть следующей архитектуры (рис.):

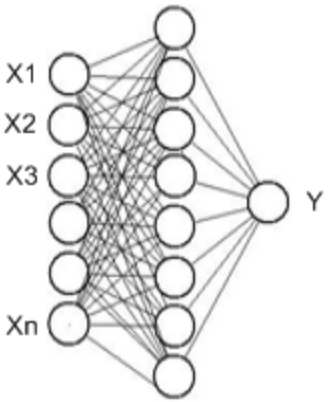

Рис. Архитектура искусственной нейронной сети GRNN

Параметры результатов обучения для выбранной архитектуры сети показаны в табл. 2.

Таблица 2

Параметры результатов обучения GRNN-сети

Параметры обучения	**Значения параметров**
Архитектура сети	GRNN 153-182-1
Средняя абсолютная разность реальных и моделируемых значений	5,15
Коэффициент корреляции расчетных и фактических значений	0,982
Отношение стандартного отклонения ошибки сети к стандартному отклонению исходных данных	0,8

Из табл. 2 можно сделать вывод, что параметры результатов обучения удовлетворительны. Коэффициент корреляции расчетных и фактических значений равен 0,982. Это говорит о хорошей сходимости модели и является наиболее важным показателем сети для решения данной задачи.

Для того чтобы оценить качество восстановления данных, в зависимости от количества исходных пропусков, из идеального массива было создано 9 отдельных массивов, с разным количеством искусственно внесенных случайных пропусков, путем сознательного удаления у некоторых наблюдений известных значений интересующих нас переменных.

Далее для каждого из массива произведена процедура восстановления данных на обученной нейронной сети. Результаты показаны в табл. 3.

Можно сделать вывод, что модель достаточно точно восстанавливает данные, если количество пропусков не превышает 5%.

Выводы. В статье проведен обзор существующих методов для заполнения пропусков в неполных данных и показана их классификация. Показаны преимущества метода, основанного на использовании нейросетевых технологий для восстановления пропущенных значений параметров в таблицах биохимических анализов пациентов с псориазом.

Предложена искусственная нейронная сеть для восстановления пропусков данных в биохимических исследованиях. Проведен точный расчет, подставляемых на место пропуска значений, и адекватность восстановления данных на искусственно введенных пропусках.

Таблица 3

Ошибка восстановления данных в зависимости от количества пропусков

Пропущено значений, %	Относительная ошибка восстановления
1%	0,012
5%	0,046
10%	0,122
15%	0,173
20%	0,193
30%	0,245
40%	0,266
50%	0,483
60%	0,591
70%	0,675

Список литературы: **1.** *Бакулев А.Л.* Псориаз как системная патология / *А.Л. Бакулев, Ю.В. Шагова, И.В. Козлова* // Саратовский научно-медицинский журнал. – 2008. – № 1 (19). – С. 13-20. **2.** *Литтл Р.Дж.А.* Статистический анализ данных с пропусками / *Р.Дж.А. Литтл, Д.Б. Рубин.* – М.: Финансы и статистика, 2001. – 254 с. **3.** *Снитюк В.Е.* Эволюционный метод восстановления пропусков в данных / *В.Е. Снитюк* // Интеллектуальный анализ информации. Межд. конф. – К. – 2006. – С. 262-271. **4**. *Красногорская Н.Н.* Применение искусственных нейронных сетей при восстановлении пропущенных гидрологических данных / *Н.Н. Красногорская, А.Н. Елизарьев, Э.В. Нафикова, Л.М. Якупова* // Промышленная экология и безопасность жизнедеятельности. – 2009. – № 1. – С. 12-16. **5.** *Васильев В.И.* Восстановление пропусков и обнаружение ошибок в эмпирических таблицах / *В.И. Васильев* // Искусственный интеллект. – 2003. – № 3. – С. 317-324. **6.** *Хайкин С.* Нейронные сеты: полный курс / *С. Хайкин.* – М.: Вильямс. – 2006. – 1104 с. **7.** *Осовский С.* Нейронные сети для обработки информации / *С. Осовский* / Пер. с польск. И.Д. Рудинского. – М.: Финансы и статистика, 2004. – 344 с. **8.** *Уоссермен Ф.* Нейрокомпьютерная техника: Теория и практика / *Ф. Уоссермен.* – М.: Мир. – 1992. – 423 с. **9.** *Рутковская Д.* Нейронные сети, генетические алгоритмы и нечеткие системы / *Д. Рутковская, М. Пилиньский, Л. Рутковский* / Пер. с польского И.Д. Рудинского. – М.: Горячая линия – Телеком, 2006. – 452 с.

Статья представлена д.т.н. проф. каф. ИКИ ХНУРЭ Авраменко В.П.

УДК 519:616-079.4:616.5

Відновлення пропущених значень параметрів в таблицях біохімічних аналізів пацієнтів з псоріазом / Висоцка Є.В., Бєловол А.Н., Кіріченко Ю.В. // Вісник НТУ "ХПІ". Тематичний випуск: Інформатика і моделювання. – Харків: НТУ "ХПІ". – 2010. – № 21. – С. 38 – 45.

У статті розглянуто підхід до відновлення пропущених значень за допомогою нейромережевих технологій. Створена нейронна мережа, що дозволяє відновлювати відсутні клініко-біохімічні показники пацієнтів з псоріазом. Проведено точний розрахунок значень,

які підставлені на місце пропуску, і адекватність відновлення даних на штучно введених пропусках. Іл.: 1. Табл.: 3. Бібліогр.: 9 назв.

Ключові слова: відновлення пропущених значень, нейромережеві технології, клініко-біохімічні показники, псоріаз.

UDC 519:616-079.4:616.5

Renewal of the skipped values of parameters in the tables of biochemical analyses of patients with psoriasis / Vysotskay E.V., Belovol A.N., Kirichenko Yu.V Herald of the National Technical University "KhPI". Subject issue: Information Science and Modelling. – Kharkov: NTU "KhPI". – 2010. – №. 21. – P. 38 – 45.

The article was considered approach to the restoration of missing values using neural network technology. Neural network was created to recover the missing clinical and biochemical parameters of patients with psoriasis. An accurate calculation, is substituted for the omission of values, and the adequacy of data recovery on an artificially imposed empty values. Figs: 1. Tabl.: 3. Refs: 9 titles.

Keywords: the restoration of missing values, neural network technology, clinical and biochemical parameters, psoriasis.

Поступила в редакцию 01.03.2010